THE THREE GAP THEOREM AND THE SPACE OF LATTICES

JENS MARKLOF AND ANDREAS STRÖMBERGSSON

Abstract. The three gap theorem (or Steinhaus conjecture) asserts that there are at most three distinct gap lengths in the fractional parts of the sequence $\alpha, 2\alpha, \ldots, N\alpha$, for any integer N and real number α. This statement was proved in the 1950s independently by various authors. Here we present a different approach using the space of two-dimensional Euclidean lattices.

Imagine we divide a cake by cutting a first wedge at an angle α, then an identical second, third, and so on as illustrated in Figure 1 (left), until the remaining piece is either of the same size as the previous, or smaller. We now have a cake comprising wedges of at most two distinct sizes: the size of the original and that of the left-over wedge. Suppose we continue cutting but insist that after each cut we rotate the knife by the same angle α as before, see Figure 1 (right). How many different sizes of cake wedges are there after N cuts? The celebrated "three gap theorem" states that for each N there will be at most three! This surprising fact was understood by number theorists in the late 1950s [6, 7, 8, 9]. Various new proofs have appeared since then, with connections to continued fractions [5, 10], Riemannian geometry [1] and elementary topology [4, App. A], as well as higher-dimensional generalisations [2, 3, 11]. Our aim here is to provide a simple proof of the three gap phenomenon by exploiting the geometry of the space of two-dimensional Euclidean lattices.

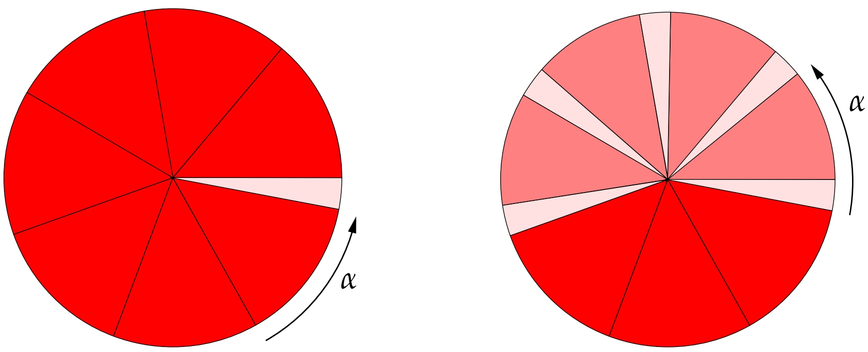

Figure 1. For each given N, there are at most three different wedge sizes.

The standard example of a Euclidean lattice in $\mathbb{R}^2$ is the square lattice $\mathbb{Z}^2$. We can generate any other Euclidean lattice $\mathcal{L}$ in $\mathbb{R}^2$ by applying a linear transformation to $\mathbb{Z}^2$. Writing points in $\mathbb{R}^2$ as row vectors $\boldsymbol{x} = (x_1, x_2)$, we have explicitly

$$\mathcal{L} = \mathbb{Z}^2 M = \{(m,n)M \mid (m,n) \in \mathbb{Z}^2\}, \tag{1}$$

Date: 14 December 2016.
2010 *Mathematics Subject Classification.* 11H06; 11J71, 52C05.
To appear in the American Mathematical Monthly.

where M is a 2×2 matrix with real coefficients. If

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \det M = ad - bc \neq 0, \tag{2}$$

then a basis of the lattice $\mathcal{L} = \mathbb{Z}^2 M$ is given by the linearly independent vectors

$$\boldsymbol{b}_1 = \boldsymbol{e}_1 M = (a, b), \qquad \boldsymbol{b}_2 = \boldsymbol{e}_2 M = (c, d), \tag{3}$$

where $\boldsymbol{e}_1 = (1, 0)$, $\boldsymbol{e}_2 = (0, 1)$ is the standard basis of $\mathbb{Z}^2$. All other bases of $\mathcal{L}$ with the same orientation can be obtained by replacing M by γM provided $\gamma \in \Gamma = \mathrm{SL}(2, \mathbb{Z})$, the group of matrices with integer coefficients and unit determinant. In the following we restrict our attention to lattices $\mathcal{L} = \mathbb{Z}^2 M$ whose basis vectors span a parallelogram of unit area. This means that $\det M = \pm 1$, and by reversing the orientation of a basis vector where necessary (this will not change the lattice), we can assume in fact that $\det M = 1$. Let us therefore denote by $G = \mathrm{SL}(2, \mathbb{R})$ the group of real matrices with unit determinant. The "modular group" $\Gamma = \mathrm{SL}(2, \mathbb{Z})$ is a discrete subgroup of G, and the space of lattices can in this way be identified with the coset space $\Gamma \backslash G = \{\Gamma g \mid g \in G\}$.

In order to translate the three gap problem into the setting of lattices, let us measure all angles in units of 360°. That is, angles are parametrized by the coset space $\mathbb{R}/\mathbb{Z} = \{x + \mathbb{Z} \mid x \in \mathbb{R}\}$ (the set of reals taken modulo one), which we can think of as the unit interval $[0, 1]$ with the endpoints 0 and 1 identified. Fix $\alpha \in \mathbb{R}/\mathbb{Z}$, and let $\xi_k = \{k\alpha\}$ be the fractional part of $k\alpha$. The quantity ξ_k represents the angular position of the kth cut. The angles of the resulting cake wedges after N cuts are precisely the gaps between the elements of the sequence $(\xi_k)_{k=1}^N$ on $\mathbb{R}/\mathbb{Z}$. These gaps are, in other words, the lengths of the N intervals that $\mathbb{R}/\mathbb{Z}$ is partitioned into by $(\xi_k)_{k=1}^N$.

The gap between ξ_k and its *next* neighbor on $\mathbb{R}/\mathbb{Z}$ (this is not necessarily the *nearest* neighbor, as the gap to the element preceding ξ_k may be the smaller one) is given by

$$s_{k,N} = \min\{(\ell - k)\alpha + n \geq 0 \mid (\ell, n) \in \mathbb{Z}^2,\ 0 < \ell \leq N,\ \ell \neq k\}. \tag{4}$$

The substitution $m = \ell - k$ yields

$$s_{k,N} = \min\{m\alpha + n \geq 0 \mid (m, n) \in \mathbb{Z}^2,\ -k < m \leq N - k,\ m \neq 0\}. \tag{5}$$

We now claim that in fact

$$s_{k,N} = \min\{m\alpha + n \geq 0 \mid (m, n) \in \mathbb{Z}^2 \setminus \{\boldsymbol{0}\},\ -k < m \leq N - k\}. \tag{6}$$

To see this, we note that the minimum in (6) is taken over a larger set than that in (5), where the additional elements correspond to $m = 0$ and $n \neq 0$. For these values $\min\{m\alpha + n \geq 0\} = 1$, which means they do not contribute to the minimum in (6). We rewrite (6) as

$$s_{k,N} = \min\{y \geq 0 \mid (x, y) \in \mathbb{Z}^2 A_1 \setminus \{\boldsymbol{0}\},\ -k < x \leq N - k\}, \tag{7}$$

with the matrix

$$A_1 = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}. \tag{8}$$

The lattice $\mathbb{Z}^2 A_1$ and $s_{k,N}$ are illustrated in Figure 2.

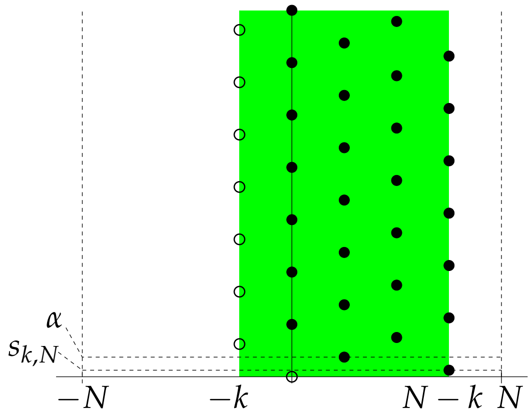

FIGURE 2. Illustration of the the expression for $s_{k,N}$ in (7) (here $N = 4$, $k = 1$).

Now take a general element $M \in G$ and $0 < t \leq 1$, and define the function F by

$$F(M,t) = \min\left\{ y \geq 0 \,\middle|\, (x,y) \in \mathbb{Z}^2 M \setminus \{\mathbf{0}\},\ -t < x \leq 1-t \right\}. \tag{9}$$

To see the connection of F with the gap $s_{k,N}$, define

$$A_N = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix} \begin{pmatrix} N^{-1} & 0 \\ 0 & N \end{pmatrix} \in G, \tag{10}$$

and note that, by rescaling the set in (7), we have

$$s_{k,N} = \frac{1}{N} \min\left\{ y \geq 0 \,\middle|\, (x,y) \in \mathbb{Z}^2 A_N \setminus \{\mathbf{0}\},\ -\frac{k}{N} < x \leq 1 - \frac{k}{N} \right\}. \tag{11}$$

Thus,

$$s_{k,N} = \frac{1}{N} F\left(A_N, \frac{k}{N} \right). \tag{12}$$

We first check F is well-defined as a function on the space of lattices $\Gamma\backslash G$ (Proposition 1), and then establish that the function $t \mapsto F(M,t)$ only takes at most three values for every fixed $M \in G$ (Proposition 2). The latter implies the three gap theorem via (12).

Proposition 1. *F is well-defined as a function $\Gamma\backslash G \times (0,1] \to \mathbb{R}_{\geq 0}$.*

Proof. Let us begin by showing that

$$\left\{ y \geq 0 \,\middle|\, (x,y) \in \mathbb{Z}^2 M \setminus \{\mathbf{0}\},\ -t < x \leq 1-t \right\} \tag{13}$$

is nonempty for every $M \in G$, $t \in (0,1]$. Let

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{14}$$

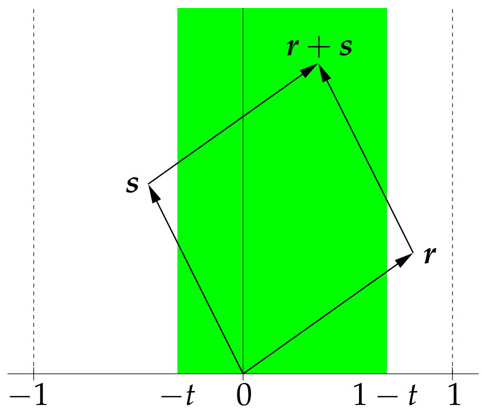

FIGURE 3. Illustration of the lattice configuration in the proof of Proposition 2.

and assume first that $a=0$. Then $c\neq 0$ and $b=-1/c$, and (13) becomes

$$\left\{ bm+dn\geq 0 \;\middle|\; (m,n)\in\mathbb{Z}^2\setminus\{\mathbf{0}\},\ -t<cn\leq 1-t \right\} \supset |b|\mathbb{N}, \tag{15}$$

which is nonempty. If $a\neq 0$, we have

$$M=\begin{pmatrix} a & b\\ c & d\end{pmatrix}=\begin{pmatrix} a & 0\\ c & a^{-1}\end{pmatrix}\begin{pmatrix} 1 & ba^{-1}\\ 0 & 1\end{pmatrix}, \tag{16}$$

and so (13) equals

$$\left\{ y+ba^{-1}x\geq 0 \;\middle|\; (x,y)\in\mathbb{Z}^2\begin{pmatrix} a & 0\\ c & a^{-1}\end{pmatrix}\setminus\{\mathbf{0}\},\ -t<x\leq 1-t \right\}. \tag{17}$$

Since $-t<x\leq 1-t$ implies $|x|\leq 1$, the set in (17) contains the set

$$\begin{aligned}(18)\quad &\left\{ y+ba^{-1}x \;\middle|\; (x,y)\in\mathbb{Z}^2\begin{pmatrix} a & 0\\ c & a^{-1}\end{pmatrix}\setminus\{\mathbf{0}\},\ -t<x\leq 1-t,\ y\geq |ba^{-1}| \right\}\\ &\qquad=\left\{ bm+dn \;\middle|\; (m,n)\in\mathbb{Z}^2\setminus\{\mathbf{0}\},\ -t<am+cn\leq 1-t,\ n\geq |b| \right\}.\end{aligned}$$

If c/a is rational, there exist $(m,n)\in\mathbb{Z}^2\setminus\{\mathbf{0}\}$ with $n\geq |b|$ such that $am+cn=0$. If c/a is irrational, then the set $\{am+cn \mid (m,n)\in\mathbb{Z}^2\setminus\{\mathbf{0}\},\ n\geq |b|\}$ is dense in $\mathbb{R}$. Therefore, in both cases, (18) is nonempty, and the minimum of (13) exists due to the discreteness of $\mathbb{Z}^2M$.

Finally, we note that $F(\,\cdot\,,t)$ is well-defined on $\Gamma\backslash G$ since $F(M,t)=F(\gamma M,t)$ for all $M\in G$, $\gamma\in\Gamma$. $\square$

The following assertion implies the classical three gap theorem; recall (12).

Proposition 2. *For every given $M\in G$, the function $t\mapsto F(M,t)$ is piecewise constant and takes at most three distinct values. If there are three values, then the third is the sum of the first and second.*

Proof. Among all points of the set $\mathcal{L}\setminus\{\mathbf{0}\}$ with $\mathcal{L}=\mathbb{Z}^2M$ in the region $\mathcal{A}=(-1,1)\times[0,\infty)$, let $\boldsymbol{r}=(r_1,r_2)$ be a point with minimal second coordinate r_2. See

Figure 3. Let us assume $r_2 > 0$ (the case $r_2 = 0$ is treated at the end of the proof). Next let $\boldsymbol{s} = (s_1, s_2)$ be a point in $\mathcal{A} \cap \mathcal{L} \setminus \mathbb{Z}\boldsymbol{r}$ with s_2 minimal. Then $s_2 \geq r_2 > 0$.

The parallelogram $\boldsymbol{0}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{r}+\boldsymbol{s}$ does not contain any other lattice points: if $\boldsymbol{u}$ were such a lattice point, then $\boldsymbol{u}$ or $\boldsymbol{r}+\boldsymbol{s}-\boldsymbol{u}$ would have second coordinate smaller than s_2, contradicting the assumed minimality of s_2. This implies that $\boldsymbol{r}, \boldsymbol{s}$ form a basis of $\mathcal{L}$.

Note that r_1 and s_1 must have opposite signs, i.e. $r_1 s_1 < 0$, since otherwise $\boldsymbol{s} - \boldsymbol{r} \in \mathcal{A}$ with a second coordinate that is smaller than s_2, contradicting the assumed minimality of s_2. It follows that, if we set $\mathcal{J}_{\boldsymbol{r}} = (0,1] \cap (-r_1, 1-r_1]$ and $\mathcal{J}_{\boldsymbol{s}} = (0,1] \cap (-s_1, 1-s_1]$, then one of these intervals is of the form $(0,q]$ and the other is of the form $(q', 1]$, for some $q, q' \in (0,1)$. Note that both intervals are nonempty since $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{A}$ by construction, and thus $|r_1|$, $|s_1| < 1$. More explicitly,

$$\mathcal{J}_{\boldsymbol{r}} = \begin{cases} (-r_1, 1] & \text{if } -1 < r_1 \leq 0 \\ (0, 1-r_1] & \text{if } 0 \leq r_1 < 1, \end{cases} \tag{19}$$

and similarly for $\mathcal{J}_{\boldsymbol{s}}$. Now in view of definition (9), we obtain

$$F(M,t) = \begin{cases} r_2 & \text{if } t \in \mathcal{J}_{\boldsymbol{r}} \\ s_2 & \text{if } t \in \mathcal{J}_{\boldsymbol{s}} \setminus \mathcal{J}_{\boldsymbol{r}} \\ r_2 + s_2 & \text{if } t \in (0,1] \setminus (\mathcal{J}_{\boldsymbol{r}} \cup \mathcal{J}_{\boldsymbol{s}}). \end{cases} \tag{20}$$

(Here the set $(0,1] \setminus (\mathcal{J}_{\boldsymbol{r}} \cup \mathcal{J}_{\boldsymbol{s}})$ may be empty.) Thus, for any fixed M, the function $F(M, \cdot\,)$ can only take one of the three values $r_2, s_2, r_2 + s_2$.

Now consider the remaining case $r_2 = 0$. Let us then also require that $\boldsymbol{r} = (r_1, r_2)$ is a primitive lattice point (primitive means that there is no lattice point on the line segment between $\boldsymbol{0}$ and $\boldsymbol{r}$) and again let $\boldsymbol{s} = (s_1, s_2)$ be a point in $\mathcal{A} \cap \mathcal{L} \setminus \mathbb{Z}\boldsymbol{r}$ with s_2 minimal (then $s_2 > 0$). If $|r_1| \leq 1/2$ then $F(M,t) = 0$ for all $t \in (0,1]$. On the other hand, if $|r_1| > 1/2$ then $F(M,t) = s_2$ for $t \in (1-|r_1|, |r_1|]$ and $F(M,t) = 0$ for all other t in $(0,1]$. $\square$

Acknowledgment. We thank both referees for their valuable feedback. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 291147. A.S. is supported by a grant from the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine, and also by the Swedish Research Council Grant 621-2011-3629.

References

[1] I. Biringer and B. Schmidt, The three gap theorem and Riemannian geometry, *Geom. Dedicata* **136** (2008) 175–190.

[2] P. Bleher, Y. Homma, L. Ji, R. Roeder and J. Shen, Nearest neighbor distances on a circle: multidimensional case, *J. Stat. Phys.* **146** (2012) 446–465.

[3] N. Chevallier, Cyclic groups and the three distance theorem, *Canad. J. Math.* **59** (2007) 503–552.

[4] A. Haynes, H. Koivusalo, J. Walton and L. Sadun, Gaps problems and frequencies of patches in cut and project sets, *Math. Proc. Camb. Philos. Soc.* **161** (2016) 65–85.

[5] N. Slater, Gaps and steps for the sequence $n\theta$ mod 1, *Proc. Camb. Phil. Soc.* **63** (1967) 1115–1123.

[6] V. Sós, On the theory of Diophantine approximations I, *Acta Math. Acad. Sci. Hungar.* **8** (1957) 461–472.

[7] V. Sós, On the distribution mod 1 of the sequence $n\alpha$, *Ann. Univ. Sci. Budapest Eötvös Sect. Math.* **1** (1958) 127–134.
[8] J. Surányi, Über die Anordnung der Vielfachen einer reellen Zahl mod 1, *Ann. Univ. Sci. Budapest Eötvös Sect. Math.* **1** (1958) 107–111.
[9] S. Świerczkowski, On successive settings of an arc on the circumference of a circle, *Fund. Math.* **46** (1959) 187–189.
[10] T. van Ravenstein, The three gap theorem (Steinhaus conjecture), *J. Austral. Math. Soc. Ser.* A **45** (1988) 360–370.
[11] S. Vijay, Eleven Euclidean distances are enough, *J. Number Theory* **128** (2008) 1655–1661.

School of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.
j.marklof@bristol.ac.uk

Department of Mathematics, Box 480, Uppsala University, SE-75106 Uppsala, Sweden
astrombe@math.uu.se